基于声学特征的语言情感识别

金 琴[1,2] 陈师哲[2] 李锡荣[2] 杨 刚[2] 许洁萍[2]
(中国人民大学数据工程与知识工程教育部重点实验室 北京 100872)[1]
(中国人民大学信息学院 北京 100872)[2]

摘 要 语音情感识别是语音处理领域中一个具有挑战性和广泛应用前景的研究课题。探索了语音情感识别中的关键问题之一:生成情感识别的有效的特征表示。从4个角度生成了语音信号中的情感特征表示:(1)低层次的声学特征,包括能量、基频、声音质量、频谱等相关的特征,以及基于这些低层次特征的统计特征;(2)倒谱声学特征根据情感相关的高斯混合模型进行距离转化而得出的特征;(3)声学特征依据声学词典进行转化而得出的特征;(4)声学特征转化为高斯超向量的特征。通过实验比较了各类特征在情感识别上的独立性能,并且尝试了将不同的特征进行融合,最后比较了不同的声学特征在几个不同语言的情感数据集上的效果(包括IEMOCAP英语情感语料库、CASIA汉语情感语料库和Berlin德语情感语料库)。在IEMOCAP数据集上,系统的正确识别率达到了71.9%,超越了之前在此数据集上报告的最好结果。

关键词 语音情感识别,声学特征,特征融合
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.9.005

Speech Emotion Recognition Based on Acoustic Features

JIN Qin[1,2] CHEN Shi-zhe[2] LI Xi-rong[2] YANG Gang[2] XU Jie-ping[2]
(Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education, Renmin University of China, Beijing 100872, China)[1]
(School of Information, Renmin University of China, Beijing 100872, China)[2]

Abstract Emotion recognition from speech is a challenging research area with wide applications. This paper explored one of the key aspects of building an emotion recognition system: generating suitable feature representation. We extracted features from four angles: (1) low-level acoustic features such as intensity, F0, jitter, shimmer, spectral contours etc. and statistical functions over these features, (2) a set of features derived from segmental cepstral-based features scored against emotion-dependent Gaussian mixture models, (3) a set of features derived from a set of low-level acoustic codewords, (4) GMM supervectors constructed by stacking the means or covariance or weights of the adapted mixture components on each utterance. We applied these features for emotion recognition independently and jointly and compared their performance within this task. We built a support vector machine (SVM) classifier based on these features. We tested the performance of these different features on some public emotion recognition corpus (including IEMOCAP corpus in English, CASIA corpus in Mandarin, and BerlinEMO-DB in Germany). On the IEMOCAP database, the four-class emotion recognition accuracy of our system is 71.9%, which outperforms the previously reported best results on this dataset.

Keywords Speech emotion recognition, Acoustic features, Feature fusion

1 引言

一般认为人类是通过语言、表情、音乐和行为等表达模式来进行情感交流的,其中包含在语音信号中的情感信息是非常重要的信息资源,因此基于语音的情感分析的研究越来越受到人们的重视。而语音情感识别是语音信息处理技术中一个非常重要的研究方向,尤其在人机交互的应用中有重要的意义和广泛的应用前景,例如在计算机交互教学[1],以及人的精神健康辅助诊断[2]等应用中都发挥着重要的作用。

要研究语音信号中所包含的情感,首先需要根据某些特性标准对语音情感做一个有效合理的分类,然后在不同类别的基础上研究特征参数的性质。本文主要集中研究语音情感

到稿日期:2014-08-12 返修日期:2014-09-13 本文受北京市自然科学基金(4142029),中国人民大学科学研究基金(中央高校基本科研业务费专项资金)(14XNLQ01)资助。

金 琴(1972—),女,博士,副教授,CCF会员,主要研究方向为音频信息处理、多媒体内容分析与理解、自然语言处理、统计机器学习,E-mail:qjin@ruc.edu.cn;陈师哲(1994—),女,主要研究方向为音频信息处理;李锡荣(1983—),男,博士,讲师,CCF会员,主要研究方向为图像检索与多媒体内容分析;杨 刚(1979—),男,博士,讲师,主要研究方向为神经网络算法;许洁萍(1966—),女,博士,副教授,CCF会员,主要研究方向为多媒体信息处理。

识别的声学特征表示；在提取低层次的帧级声学特征的基础上，通过4种不同的转化操作生成不同的衍生特征。这4种转化包括：1)传统的统计函数变换，包括极值、均值、方差等；2)将帧级声学特征与情感相关的高斯混合模型进行距离计算而转化的特征；3)通过数据驱动方法得到的声学码字，并以这些码字进行转化的特征；4)通过高斯混合模型转化生成的超向量特征表示。本文也探索了将不同特征进行融合，包括前期在特征级别的融合以及后期在分类结果级别的融合。

本文第2节简要介绍了语音情感识别的相关工作；第3节详细介绍了低层声学特征以及通过4种变化衍生的特征；第4节介绍了实验所用的3个数据集；第5节阐述了具体的实验设置以及实验结果；最后对本文的研究工作进行总结和展望。

2 相关工作

语音情感描述方式大致可以分为离散情感类别和连续情感维度两种形式。前者将情感描述成离散的、情感类别标签的形式，如高兴、悲伤等，属于分类问题；后者则将情感状态描述为多维情感空间的点，每个维度对应着情感空间的一个心理学属性，如表示情感激烈程度的激活度属性。其因用连续的实数对每一维度进行描述，也被称为连续情感描述，一般被建模为标准的预测或拟合问题。这两种形式都具有各自表达情感的优缺点，离散情感描述简洁、易懂、容易着手，但是其单一有限的情感描述能力无法满足对自发情感的描述；连续情感描述拥有无限的情感描述能力，但将主观情感转化为客观数值的过程是繁重且无法保证质量的。当前，离散情感分类的研究比连续情感维度的研究发展更为繁荣。在目前离散语音情感分类研究中，常用的情感分为8类情感模型(高兴、期望、愤怒、厌恶、悲伤、惊奇、恐惧、赞同)或4类情感模型(喜、怒、惊、悲)。

对于离散情感识别而言，其系统的识别准确率是与两个要素紧密相关的：特征表示和分类器。在语音情感识别研究领域，研究人员们已经尝试了各种不同的分类技术，包括高斯混合模型(Gaussian Mixture Model，GMM)、隐马尔可夫模型(Hidden Markov Model，HMM)、K-近邻(K-nearest neighbor，KNN)、人工神经网络(Artificial Neural Networks，ANN)、支持向量机(Support Vector Machine，SVM)等[4-9]。其中支持向量机被认为是对不同的模式识别问题可以得到比其他的传统分类技术更好、更泛化的性能的方法。

语音情感的变化通常可以体现为语音特征参数的变化。例如高兴时，通常是语速较快，音量较大；悲伤时，通常是语速缓慢，音量较小。声学特征(Acoustic Feature)是语音情感识别系统使用的最主要特征。研究人员们探索了很多不同的声学特征，包括与基频(Pitch)、能量(Energy)、语速(Speech Rate)、共振峰等相关的韵律特征[10,11]以及频谱相关的特征，例如Mel-Frequency Cepstral Coefficients(MFCC)和Perceptual Linear Prediction(PLP)等[8]。其中MFCC是目前使用最广泛的语音特征之一，具有计算简单、区分能力好等突出的优点。这些特征大部分是帧级的特征。基于帧级特征的统计特征(例如均值、方差、范围等)也被广泛应用于语音情感识别系统中[5]。近年来，根据深度神经元网络学习得到的特征也在语音情感识别任务中取得了很好的性能[12]。但是基于深度神经元网络的特征的学习需要大量的训练数据，本文没有直接与其进行比较。本文工作主要集中在生成语音情感识别的有效声学特征表示。

3 声学特征

本文首先对每个语音句子提取了帧级的低层次声学特征，然后再整体地或局部地对这些低层次基础特征进行转化操作。

3.1 低层次基础声学特征

首先利用OpenSMILE工具[13]进行低层次特征提取，参考了Interspeech 2010年泛语言学挑战赛(Paralinguistic Challenge)中广泛使用的特征提取配置文件“emobase2010.conf”[14]。表1列出了本文实验中所抽取的低层次基础声学特征。其中基频特征和声音质量特征是用40ms的帧窗和10ms的帧移抽取，倒谱类的特征是用25ms的帧窗和10ms的帧移抽取。

表1 低层次基础声学特征

FEATURES	DESCRIPTION
Loudness+Delta	Then loudness as the normalized intensity raised to a power of 0.3
F0final+Delta	The smoothed fundamental frequency contour
F0finEnv+Delta	The envelope of the smoothed fundamental frequency contour
jitterLocal+Delta	The local(frame-to-frame) Jitter (pitch period length deviations)
jitterDDP+Delta	The differential frame-to-frame Jitter (the 'Jitter of the Jitter')
shimmerLocal+Delta	The local(frame-to-frame) Shimmer (amplitude deviations between pitch periods)
Voicing final+Delta	The voicing probability of the final fundamental frequency candidate.
MFCC-related	MFCCs(15)+logMelFreqBand(8)

3.2 统计函数转化的声学特征

在基础声学特征上应用了21个不同的统计函数，将每个句子的一组时长不等的基础声学特征转化为定长的静态特征。这些统计函数包括最大最小值、均值、时长、方差等。关于这些统计函数的具体描述可以参考文献[13]。

3.3 模型转化的声学特征

这一转化的目的也是将基于分段的时长不等的倒谱特征转化成一组新的定长的静态特征。但是新的特征要保持情感区分的信息，而不只是简单地在句子级别上进行统计计算(如3.2节中统计函数转化的声学特征)。

首先，基于倒谱特征为每类情感分别训练一个有5个高斯分量的高斯混合模型。整个高斯混合模型或者其中的高斯分量都可以被看作是情感相关的模型。通过计算倒谱特征与情感相关模型之间的匹配度或者距离来进行转化。转化后的特征包含3个维度(p,h,α)[15]。其中p是归一化的帧级倒谱特征与情感模型匹配概率得分的平均值，h是匹配概率得分高的比率，α是在帧级特征分布为Dirichlet分布的假设前提下转化生成的高级特征。将这些新特征称作模型转化的倒谱特征(M-Cepstrum)。

3.4 码字转化的声学特征

码本技术是在文本分类(bag-of-words词袋)以及图像分类(bag-of-visual words视觉词袋)等任务中常用的技术。类似的音频词袋(bag-of-audio words)的方法也被成功地应用到

多媒体事件检测等任务中[16]。其基本思想就是对于一段音频上的基础倒谱特征，通过统计其在码本中每个码字上的分布，将其转化为维度为码本大小的新特征。本文首先使用 K-近邻(K-means)聚类算法产生一个声学码本，然后将每个句子表示成其基础声学特征在每个码字上的分布：

$$d_i=(d_{i,1},\cdots,d_{i,K})$$

其中，$d_{i,j}$ 代表第 i 个句子中属于第 j 个码字的帧的数目的加权平均。K 代表码本的大小即码字的个数，这个参数可以通过在开发集数据上学习调整到最优。

3.5 高斯超向量特征

近年来，高斯超向量在话者识别的任务中有很成功的应用[17]。高斯超向量通常是通过拼接高斯混合模型中的均值或协方差或权值而生成的。首先在随机抽取的包含所有情感的数据上训练出一个通用背景高斯混合模型(称为 GMM-UBM)：

$$g(X)=\sum_{i=1}^{M}\lambda_i N(X;U_i,\Sigma_i)$$

其中，λ_i 是权值，$N(U_i,\Sigma_i)$ 是单个高斯，U_i 和 Σ_i 是高斯的均值和协方差。假定协方差 Σ_i 是对角阵，对于每个句子，可以通过 MAP(Maximum A Posterior)适应生成一个对应的高斯混合模型(Adapted GMM)；然后拼接这个高斯混合模型的均值 U_i 或对角阵 Σ_i 或权值 λ_i 来产生不同的超向量，图 1 示出了拼接均值产生超向量的过程。可以把高斯超向量看作是由低层次声学特征转化的高维特征，并将其作为 SVM 分类器的输入特征。

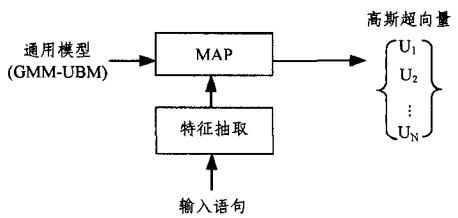

图 1 高斯超向量特征的生成图示

4 数据集描述

分别在 IEMOCAP 英语情感语料库、CASIA 汉语情感语料库和 Berlin 德语情感语料库中进行了实验。下面将对这 3 个数据集逐一介绍。

4.1 IEMOCAP 英语情感数据集

IEMOCAP 是由南加利福利亚大学录制的情感数据库，包含约 12 小时的视听数据，即视频、音频和语音文本、面部表情[18]。10 名专业演员(5 男 5 女)在有台词或即兴的场景下，特意引导出情感表达。之后，人工将每一段对话切分成单句，每一句话至少由 3 个标注员进行类别标注(如高兴、生气等)，此外也对 Valence、Activation、Dominance 3 个维度进行了标注。本文的实验仅考虑无较大争议的分类标注结果。为了平衡不同情感类别的数据，将高兴(happy)和兴奋(exciting)合并成高兴类别。由高兴、生气、悲伤和中性最终构成了 4 类情感识别数据库。表 2 展示了每一类情感的语句个数。

表 2 IEMOCAP 数据集中每个情感类别语句的数量

生气	高兴	悲伤	中性	总计
1103	1636	1084	1708	5531

4.2 CASIA 汉语情感语料库

CASIA 是由中国科学院自动化研究所录制的[18]。语料设计包含 6 类不同情感：高兴、悲哀、生气、惊吓、难过、中性。每种情感有 50 句语料，由 4 位录音人(2 男 2 女)在纯净录音环境中(信噪比约为 35dB)对 50 句语料赋予不同的情感演绎而得到。语音信号采用 16kHz 采样以及 16bit 量化。经过听辨筛选，最终保留 1200 句语音样例。

4.3 柏林 EMO-DB 德语情感语音库

EMO-DB 是由柏林工业大学录制的德语情感语音库[20]，由 10 名演员(5 男 5 女)对 10 个语句(5 长 5 短)进行 7 种感情(高兴、生气、焦虑、害怕、无聊、厌恶和中性)的演绎而得到，共包含 535 句语料。语音信号同样采用 16kHz 采样以及 16bit 量化。语料文本的选取遵从语义中性、无情感倾向的原则，且为日常口语化风格，无过多的书面语修饰。语音的录制在专业录音室中完成，要求演员在演绎某个特定情感前通过回忆自身真实经历或体验进行情绪的酝酿，来增强情绪的真实感。经过 20 个参与者(10 男 10 女)的听辨实验，得到 84.3%的听辨正确率。

5 实验

本文中所有的实验都是关于语音情感分类的。实验的评测标准是识别准确率，即：

$$\text{准确率}=\frac{\text{正确识别的测试样例数}}{\text{总的测试样例数}}$$

5.1 实验设置

本文中所有的实验都是在 10 组交叉验证(10-fold cross validation)的模式下完成。

在后文中出现的不同声学特征及其衍生特征的名称缩写及描述如下所示：

- Cepstrum：在帧级的倒谱声学特征上应用统计函数得到语句级别的倒谱特征。
- ACO：在帧级的基础声学特征(除去倒谱 Cepstrum 特征)上应用统计函数得到语句级别的特征。
- M-Cepstrum：倒谱声学特征根据情感相关的高斯混合模型进行距离转化而得出的特征。
- BoW：利用声学码本转化的特征。ACO-BoW 是指对帧级基础声学特征(除倒谱特征)进行码本转化而得到的特征，Cepstrum-BoW 是指对倒谱声学特征进行码本转化而得到的特征。
- GSV：通过拼接 MAP-自适应得到的高斯混合模型的均值或协方差或权重而形成的高斯超级向量。在实验中，生成的 GSV 特征是基于帧级的倒谱声学特征而得到的。
- +：特征的拼接操作。如：ACO+Cepstrum 是指拼接 ACO 和 Cepstrum 特征。

实验中，首先利用 SVM[21]分类器在 IEMOCAP 上比较了上述每组特征的效果，并尝试了前期特征融合和后期分类结果融合，然后将较好的特征及组合推广到 CASIA 和 EMO-DB 不同语言的数据集中来验证其鲁棒性和可迁移性。

分类器是情感识别系统中最重要的部分之一。在众多的分类器中，支持向量机 SVM 在多种不同应用中都被认为是最有效的分类器之一，而且比神经网络更易于使用。通常，径向基核函数(RBF-SVM)是最基础的选择。C 和 γ 是在 RBF-

SVM 中需要调节的两个参数，C 控制调整训练的错误和最大边界，γ 调整核的宽度。通常使用网格搜索进行交叉验证，对 C 和 γ 进行调优。在实验中，使用数据挖掘工具 Weka[22] 中的网格搜索算法来调整 C 和 γ，其中 C 的范围是 2^{-1} 到 2^{4}，γ 的范围是 2^{-7} 到 2^{-2} 呈指数增长。

5.2 IEMOCAP 实验结果

表 3 列出了每一组特征在 IEMOCAP 数据集上的分类效果。基于协方差的高斯超向量取得了在单一特征中的最高准确率 67.8%，超过了此前文献[15]中报告的声学特征在 IEMOCAP 4 类情感分类的准确率。码本转化(BoW)选择的聚类码本大小为 4096，其在倒谱声学特征(Cepstrum)中效果良好，但是在基础声学特征(ACO)中的效果却大相径庭。这是因为对于尤其是描述语音质量的特征来说，其包含了长时信息，仅仅有小部分的帧有非零值；而对于频谱特征蕴含的短时信息，则可以很好地通过码本转换体现。

表 3 单组特征在 IEMOCAP 数据集上的分类准确率

特征集	特征维数	准确率
ACO	616	62.5%
Cepstrum	966	65.1%
M-Cepstrum	52	59.8%
ACO-BoW	4096	52.2%
Cepstrum-BoW	4096	66.4%
(ACO+Cepstrum)-BoW	4096	63.4%
GSV-mean	2944	67.3%
GSV-cov	**2944**	**67.8%**
GSV-weight	64	56.6%

在前期融合中，对特征集合中所有特征的两两组合进行了实验，即不同特征的简单拼接。表 4 展示了 Top-10 的特征组合及其分类准确率。从中发现 ACO 与其他每一个单一特征的融合都对分类性能有所提高，因此考虑到 ACO 和其他短时信息特征的互补性，把 ACO 和其他频谱转化得到的特征进行 3 种特征的融合。如表 5 所列，ACO、GSV-mean 和 GSV-cov 融合取得了最高的分类准确率 71.9%，远远超出了此前在 IEMOCAP 5 类情感识别中的最好结果[15]。

表 4 Top10 两组特征前期融合的分类准确率

特征集	准确率
Cepstrum-BoW+GSV-mean	**71.00%**
Cepstrum-BoW+GSV-cov	70.90%
Cepstrum+Cepstrum-BoW	70.60%
Cepstrum+GSV-mean	70.40%
Cepstrum+GSV-cov	70.20%
ACO+Cepstrum-BoW	70.10%
M-Cepstrum+GSV-cov	69.70%
ACO+GSV-cov	69.60%
GSV-mean+GSV-cov	69.40%
ACO+GSV-mean	69.10%

表 5 3 组特征前期融合的分类准确率

特征集	准确率
ACO+Cepstrum-BoW+GSV-mean	71.6%
ACO+Cepstrum-BoW+GSV-cov	71.8%
ACO+Cepstrum+Cepstrum-BoW	71.4%
ACO+Cepstrum+GSV-mean	71.2%
ACO+Cepstrum+GSV-cov	70.3%
ACO+M-Cepstrum+GSV-cov	70.3%
ACO+GSV-mean+GSV-cov	**71.9%**

对于后期融合，采用简单的线性加权融合。利用 FoCal 工具[23]在训练集中进行权重的取优，最后将得到的权重参数应用于测试集。表 6 列出了在决策层面上的后期融合的情感分类准确率。实验结果显示后期融合效果没有前期融合好，很有可能是由于投票结果值的离散性使其并不适用于后期融合。

表 6 两组特征后期融合的情感分类准确率

特征集	准确率
Cepstrum+GSV-cov	65.5%
Cepstrum-BoW+GSV-mean	67.0%
Cepstrum-BoW+GSV-cov	67.1%
Cepstrum+Cepstrum-BoW	67.0%
Cepstrum+GSV-mean	65.4%
ACO+Cepstrum-BoW	66.9%
M-Cepstrum+GSV-cov	**68.0%**
ACO+GSV-cov	67.9%
GSV-mean+GSV-cov	67.5%
ACO+GSV-mean	67.3%

5.3 CASIA 实验结果

使用在 IEMOCAP 中效果较好的单个特征集合在 CASIA 数据集上进行测试。结果如表 7 所列，其中，对于 Cepstrum-BoW 特征，经过交叉验证当聚类码本大小为 2048 时效果最好，这是由于 CASIA 数据集的数据量比 IEMOCAP 数据集小。在中文语音环境下，Cepstrum-BoW 和 GSV-mean 仍然保持了很好的迁移性。由于后期融合效果不明显，只进行了前期的两类和三类特征集融合，实验结果如表 8 所列。

表 7 单组特征在 CASIA 数据集上的情感分类准确率

特征集	特征维数	准确率
ACO	616	72.6%
Cepstrum	966	82.8%
Cepstrum-BoW	**2048**	**85.1%**
GSV-mean	2944	83.0%
GSV-cov	2944	77.0%

表 8 多组特征前期融合在 CASIA 数据集上的情感分类准确率

特征集	准确率
ACO+Cepstrum	84.7%
ACO+Cepstrum-BoW	85.3%
ACO+GSV-mean	85.1%
Cepstrum-BoW+GSV-mean	86%
Cepstrum+Cepstrum-BoW	**87.2%**
ACO+Cepstrum-BoW+GSV-mean	86.7%
ACO+Cepstrum+Cepstrum-BoW	**87.8%**

5.4 EMO-DB 实验结果

同样的配置在 EMO-DB 数据集上的实验结果如表 9 和表 10 所列，Cepstrum-BoW 的聚类码本大小为 1024 时效果最好，这也与数据集大小相关，EMO-DB 的数据集大小约为 CASIA 的一半。

表 9 单组特征在 EMO-DB 数据集上的分类准确率

特征集	特征维数	准确率
ACO	616	80.9%
Cepstrum	**966**	**86.6%**
Cepstrum-BoW	1024	84.7%
GSV-mean	2944	86.4%
GSV-cov	2944	82.1%

表 10 多组特征前期融合在 EMO-DB 数据集上的分类准确率

特征集	准确率
ACO+Cepstrum	87.8%
ACO+Cepstrum-BoW	88.4%
ACO+GSV-mean	88.8%
Cepstrum+GSV-mean	89.5%
Cepstrum+Cepstrum-BoW	**89.7%**
ACO+Cepstrum+Cepstrum-BoW	90.1%
ACO+Cepstrum+GSV-mean	**90.3%**

从表中可以看出，Cepstrum-BoW 和 GSV-mean 对不同语言的鲁棒性较高，但是受数据集大小影响较大；数据集减小后，其效果甚至不如直接对每帧特征的统计函数值。但是，ACO 特征与经过转化后的倒谱声学特征的互补信息更多，融合效果显著。

结束语 本文在 IEMOCAP 英语情感语料库、CASIA 汉语情感语料库和 Berlin 德语情感语料库等 3 种不同语言、不同大小的数据集上进行了语音情感识别实验，主要工作是抽取不同的声学特征以及对帧级别特征的多种转化，包括模型转换、码本转换、高斯超级向量。数据集的规模对于数据驱动的码本特征和高斯超级向量有较大影响，但整体来说，转化后的特征分类效果更优，且与原始统计函数的特征具有互补性。在每个数据集中通过特征融合，都大大提高了识别准确率。在 IEMOCAP 数据集上，系统的识别准确率达到了 71.9%，超越了之前在此数据集上报告的最好结果。

未来的工作会致力于探索其他不同类型的特征转化方法，如利用深度神经网络进行特征学习以及在后期更加有效的融合模式。

参 考 文 献

[1] Litman D, Forbes K. Recognizing emotions from student speech in tutoring dialogues[C]// Proceeding of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). 2003: 25-30

[2] France D J, Shiavi R G, Silverman S, et al. Acoustical properties of speech as indicators of depression and suicidal risk [J]. IEEE Trans. on Biomedical Engineering, 2000, 47(7): 829-837

[3] Yang N, Muraleedharan R, Kohl J, et al. Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion[C]// Proceedings of the 4th IEEE workshop on Spoken Language Technology (SLT), 2012. Miami, Florida, 2012: 455-460

[4] Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture[C]// Proceedings of the ICASSP. 2004, 1: 577-580

[5] Ayadi M, Kamel M, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern Recognition, 2011, 44(3): 572-587

[6] Zeng Z, Pantic M, Rosiman G I, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009, 31(1): 39-58

[7] Kockmann M, Burget L, Cemocky J. Application of speaker and language independent state-of-the-art techniques for emotion recognition[J]. Speech Communication, 2011, 53(9): 1172-1185

[8] Chen L, Mao X, Xue Y-L, et al. Speech Emotion Recognition: Features and Classification Models[J]. Digital Signal Processing, 2012, 22(6): 1154-1160

[9] Zhang B Y, Yu J Q, Tang J F, et al. Movie background music classification foremotion [J]. Computer Science, 2013, 40(12): 37-40, 74

[10] Schuller B, Reiter S, Mueller R, et al. Speaker-independent speech emotion recognition by ensemble classification[C]// Proceedings of IEEE International Conference on Multimedia and Expo(ICME). Amsterdam, Netherlands, 2005: 864-867

[11] Pao T L, Chen Y T, Ye J H, et al. Mandarin Emotional Speech Recognition based on SVM and NN[C]// Proceedings of International Conference on Patter Recognition (ICPR). 2006, 1: 1096-1100

[12] Lee H, Largman Y, Pham P, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks [C]// Proceedings of Advances in Neural Information Processing Systems(NIPS). 2009: 1-9

[13] Eyben F, Wollmer M, Schuller B. OpenSMILE-The Munich Versatile and Fast Open-Source Audio Feature Extractor[C]// Proceedings of ACM Multimedia (MM). Florence, Italy, 2010: 1459-1462

[14] Schuller B, Batliner A, Steidl S, et al. Recognizing Realistic Emotions and Affect in Speech: State of the Art and Lessons Leant from the First Challenge[J]. Speech Communication, 2011, 53 (10): 1062-1087

[15] Rozgic V, Ananthakrishnan S, Saleem S, et al. Emotion Recognition using Acoustic and Lexical Features[C]// Proceedings of INTERSPEECH 2012. September Portland, 2012

[16] Lee K, Ellis D P W. Audio-Based Semantic Concept Classification for Consumer Video[J]. IEEE Trans. Audio, Speech, and Language Processing, 2010, 18(6): 1406-1416

[17] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification[J]. IEEE Signal Processing Letters, 2006: 308-311

[18] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Journal of Language Resources and Evaluation, 2008, 42(4): 335-359

[19] Data collected by the speech group at National Key Laboratory of Pattern Recognition[OL]. http://www.datatang.com/data/39277

[20] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]// Proceedings of INTERSPEECH 2005. Lisbon, 2005: 1517-1520

[21] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification[OL]. 2010. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[22] Witten I H, Frank E, Trigg L E, et al. Weka: Practical machine learning tools and techniques with Java implementations[OL]. http://www.cs.waikato.ac.nz/~eibe/pubs/99IHW-EF-LT-MH-GH-SJC-Tools-Java.pdf

[23] Brummer N. FoCal-II: Toolkit for calibration of multiclass recognition scores[OL]. https://sites.google.com/site/nikobrummer/focal